A Further Note on
Score Tests for Regression Models

Chan-Fu Chen

University of Minnesota
School of Statistics

Technical Report No. 433

April 1984

Academia Sinica / University of Minnesota
School of Statistics
Department of Applied Statistics
St. Paul, Minnesota

SUMMARY

The robustness nature of the score tests for the null hypothesis of no covariate-effect is further explored in the context of the regression analysis. It is shown that the resulting test statistics share the common quadratic form in terms of projection which do not depend on the form of the effect specified under the alternative hypothesis.

KEY WORDS: Score Tests; The Exponential Family; The Location Family; Locally Asymptotically Most Powerful; Robustness of Optimality.

1. INTRODUCTION

The score tests, which use the "efficient scores" as the test criteria, were first introduced by Rao (1948) for testing the homogeneity of parallel samples. They also appear in the LaGrange forms (Aitchinson and Silvey, 1958), and are optimal $c(\alpha)$ tests of Neyman (1959). The score tests are locally asymptotically most powerful (cf. Bühler and Puri, 1966; Moran, 1970), and, as homogeneity tests, "robust" in the sense that their optimality remains true whenever the form of the compounded distribution assumed for the data under the alternative hypothesis, provided that its third moment is zero (Moran, 1973)--a property termed as "robustness of optimality" by Neyman and Scott (1966).

Recently, this robustness nature of the score tests for testing either trends in means or heteroscedasticity in variances in the regression context has attracted increasing attention in statistical literature (e.g., Tarone and Gart, 1980; Tarone, 1982; Breusch and Pagan, 1979; and Cook and Weisberg, 1983). In his general approach, Chen (1983) shows that the score tests for the null model of constant mean response remain the same, regardless of the functional form chosen for the effect of covariates (or, explanatory variables) under the alternative.

The purpose of the present note is to broaden the scope of applications of the score tests and their robustness nature discussed by Chen (1983) for the simple null model to more intermediate problems focusing primarily on the effect of a certain given subset of covariates, say, $\underset{\sim}{Z}$, on the mean μ of the response variable y. The regression model considered here is

$$\theta \equiv \theta(\mu) = \text{something else} + f(\underset{\sim}{\beta}_1 \underset{\sim}{Z}') \ ,$$

where θ is a given transformation of μ, and f, which represents the functional form of the $\underset{\sim}{Z}$-effect on θ, is monotonic and twice differentiable. For instance, f may be (a) linear: $f(t) = t$, (b) multiplicative or exponential: $f(t) = ce^t$, or (c) logistic: $f(t) = e^t/(1+e^t)$. As in the null model case, we show that the score test for the null hypothesis of no $\underset{\sim}{Z}$-effect, i.e., H_0: $\underset{\sim}{\beta}_1 = \underset{\sim}{0}$, remains unchanged whenever the form chosen for f under the alternative H_1: $\underset{\sim}{\beta}_1 \neq \underset{\sim}{0}$. This unique feature of the score tests is not shared by their asymptotically equivalent counterparts, the likelihood ratio tests and the tests based on the maximum likelihood estimates; for both the form of f under H_1 is crucial and indispensable. Therefore, the score tests should be favored as the asymptotic optimal tests for checking whether a set of covariates $\underset{\sim}{Z}$ should be included in the regression <u>whenever</u> the precise knowledge of how $\underset{\sim}{Z}$ might affect the mean response under H_1 is lacking. Due to their common quadratic form in terms of projection, the resulting test statistics can be easily understood through the notion of least squares fitting.

2. THE SCORE TESTS

Let $L(\underset{\sim}{\beta})$ be the loglikelihood function of $\underset{\sim}{\beta} = (\underset{\sim}{\beta}_0, \underset{\sim}{\beta}_1)$ and $\hat{\underset{\sim}{\beta}} = (\hat{\underset{\sim}{\beta}}_0, \underset{\sim}{0})$ be the maximum likelihood estimate (MLE) of $\underset{\sim}{\beta}$ under the null hypothesis H_0: $\underset{\sim}{\beta}_1 = \underset{\sim}{0}$. The score-test statistic S for H_0 then can be expressed as

$$S = VC^{-1}V' \,, \tag{2.1}$$

where V and C, the total score and the expected Fisher information matrix of $\underset{\sim}{\beta}$ at $\underset{\sim}{\beta} = \hat{\underset{\sim}{\beta}}$, respectively, are defined as

$$V \equiv [\partial L/\partial \underset{\sim}{\beta}]_{\hat{\underset{\sim}{\beta}}}$$

$$= (\underset{\sim}{0}, V_1) \quad , \qquad V_1 = [\partial L/\partial \underset{\sim}{\beta}_1]_{\hat{\underset{\sim}{\beta}}} \quad ,$$

and

$$C = [-E(\partial^2 L/\partial \underset{\sim}{\beta}' \, \partial \underset{\sim}{\beta})]_{\hat{\underset{\sim}{\beta}}} \; .$$

Note that the form (2.1) is unchanged even if some nuisance parameter or parameters $\underset{\sim}{\phi}$ of finite dimension, such as the scale parameter of the exponential family, which satisfy $E(\partial^2 L/\partial \underset{\sim}{\beta}' \, \partial \underset{\sim}{\phi}) = 0$, are introduced to the problem. In this case, the unknown $\underset{\sim}{\phi}$ appearing in S will be replaced by the corresponding MLE $\hat{\underset{\sim}{\phi}}$ under H_0.

For various applications of score tests in the context of regression analysis, see, e.g., references cited by the recent papers such as Breusch and Pagan (1980), Pregibon (1982), Chen (1983) and Weisberg (1983).

3. THE GENERAL RESULTS

Let $h(y|\mu)$ be the probability density (distribution for the discrete case) function of a response variable y given its mean μ, and

$$s(\mu) \equiv \partial \ell n h/\partial \mu \tag{3.1}$$

the score function of μ. It is assumed that h satisfies the following regularity conditions:

(i) $E(s(\mu)|\mu) = 0$,

(ii) $d^2(\mu) \equiv -E(\partial^2 \ell n\, h/\partial \mu^2) = E(\partial \ell n\, h/\partial \mu \cdot \partial \ell n\, h/\partial \mu) = \mathrm{Var}(s(\mu))$.

Let $\underset{\sim}{X} = (\underset{\sim}{W}, \underset{\sim}{Z})$, $\underset{\sim}{W}$: $1 \times \ell$ and $\underset{\sim}{Z}$: $1 \times k$, be the vector of $\ell+k$ covariates or explanatory variables. The first component of $\underset{\sim}{W}$ is the constant one.

Denote $(\underset{\sim}{y}',X) = (\underset{\sim}{y}',W,Z)$: $n \times (1+\ell+k)$ the data matrix of n $(n > \ell+k)$ independent observations, $(y_i,\underset{\sim}{W}_i,\underset{\sim}{Z}_i)$, $i=1,\ldots,n$, with rank $(X) = \ell+k$; that is, X is of full rank. For each given $(\underset{\sim}{W}_i,\underset{\sim}{Z}_i)$, y_i is assumed to arise from $h(y|\mu_i)$, where μ_i takes this form

$$\mu_i = g(\underset{\sim}{\beta}_0\underset{\sim}{W}_i' + f(\underset{\sim}{\beta}_1\underset{\sim}{Z}_i')) \ , \tag{3.2}$$

where g, $f \varepsilon F \equiv$ {monotonic and twice differentiable functions on R}. Here the function g, whose inverse is referred to as the "link function" in the generalized linear models (Nelder and Wedderburn, 1972), relates the mean to covariates $(\underset{\sim}{W},\underset{\sim}{Z})$ via $\theta \equiv g^{-1}(\mu)$. The function f represents the form of the effect of $\underset{\sim}{Z}$ on θ. (For simplicity, the effect of $\underset{\sim}{W}$ is assumed linear.) Then, the null hypothesis of no $\underset{\sim}{Z}$-effect is equivalent to

$$H_0: \underset{\sim}{\beta}_1 = \underset{\sim}{0} \ ;$$

that is, $f(\underset{\sim}{\beta}_1\underset{\sim}{Z}') =$ constant for all $\underset{\sim}{Z}$. Without loss of generality, we shall assume $f(0) = 0$ in the following discussion.

Let $\hat{\underset{\sim}{\beta}} = (\hat{\underset{\sim}{\beta}}_0,\underset{\sim}{0})$ be the MLE of $\underset{\sim}{\beta}$ under H_0, and

$$\hat{\mu}_i = g(\hat{\underset{\sim}{\beta}}_0\underset{\sim}{W}_i') \quad , \quad i=1,\ldots,n \ ,$$

the resulting estimate of μ_i. With the loglikelihood function of $\underset{\sim}{\beta}$ being $L(\underset{\sim}{\beta}) = \Sigma \ \ell n \, h(y_i|\mu_i)$, where μ_i is given by (3.2), one can easily verify that, under the regularity conditions (i) and (ii), V and C of (2.1) become, respectively,

$$V = \underset{\sim}{s} \ G(W, \ aZ)$$

and

$$C = (W,\ aZ)'\ GD^2G(W,aZ)\ ,$$

where

$$\underset{\sim}{s} = (s(\hat{\mu}_1),\ldots,s(\hat{\mu}_n))\ ,$$

$$G = \mathrm{diag}(\dot{g}(\hat{\underset{\sim}{\beta}}_1\underset{\sim}{W}_1'),\ldots,\dot{g}(\hat{\underset{\sim}{\beta}}_1\underset{\sim}{W}_n'))\ , \tag{3.3}$$

$$D^2 = \mathrm{diag}(d^2(\hat{\mu}_1),\ldots,d^2(\hat{\mu}_n))\ ,$$

and

$$a = \dot{f}(0) \neq 0.$$

For any matrix A of full rank, denote R(A) the range space of A and $P_A = A(A'A)^{-1}A'$, the projection matrix on R(A). Then the score-test statistic S of (2.1) for H_0 becomes

$$S = \underset{\sim}{s}G(W,aZ)\{(W,aZ)'GD^2G(W,aZ)\}^{-1}(W,aZ)'G\underset{\sim}{s}'$$

$$= \underset{\sim}{e}\ P_{X*}\ \underset{\sim}{e}'\ , \tag{3.4}$$

where

$$X^* = (W^*,Z^*) \equiv DG(W,Z) \tag{3.5}$$

and

$$\underset{\sim}{e} = \underset{\sim}{s}\ D^{-1} \tag{3.6}$$

is the standardized score vector evaluated at H_0. That is, S can be expressed in terms of the square length of the projection of the standardized score vector $\underset{\sim}{e}$ on the range space R(X*) of X*. The only quantity, $a = \dot{f}(0)$, in V and C, which depends on f, disappears in the final form of S because R(aZ) = R(Z). Further note that, since $\hat{\underset{\sim}{\beta}}$ is the MLE of $\underset{\sim}{\beta}$ under H_0,

$$\underset{\sim}{e}\, W^* = \underset{\sim}{s}\, GW = [\partial L/\partial \underset{\sim}{\beta}_0]_{\hat{\underset{\sim}{\beta}}} = \underset{\sim}{0}\ .$$

Hence, S of (3.4) can be re-expressed in terms of the projection on $R(Z^*/W^*)$ as

$$S = \underset{\sim}{e}\, P_{Z^*/W^*}\, \underset{\sim}{e}' \tag{3.7}$$

with the matrix $Z^*/W^* \equiv (I - P_{W^*})Z^*$.

Assume that $\hat{\underset{\sim}{\beta}}_0$ is a consistent estimate of $\underset{\sim}{\beta}_0$. Then, based on (3.7), one can verify that S is asymptotically distributed as X^2 with k degrees of freedom under H_0, provided that, when dealing with nonnormal cases, Eicker's necessary and sufficient condition (Eicker, 1965, Theorem 3.1, Condition (I)) for asymptotic normality of the least squares estimates holds; that is,

$$\max_i\ (P_{Z^*/W^*})_{ii} \to 0, \quad \text{as } n \to \infty,$$

where $(P_{Z^*/W^*})_{ii}$ is the ith diagonal element of P_{Z^*/W^*}.

The resulting score test based on S is asymptotically most powerful in the local sense as $\underset{\sim}{\beta}_1$ approaching $\underset{\sim}{0}$ at the rate $n^{-\frac{1}{2}}$ under H_1 (Bühler and Puri, 1966, Theorem 3). Since S does not depend on f, this asymptotic optimality is "robust" with respect to the class of forms, F, for the $\underset{\sim}{Z}$-effect under H_1.

By the fact that the score vector V is the first derivative of the loglikelihood $L(\underset{\sim}{\beta})$ evaluated at $\underset{\sim}{\beta} = \hat{\underset{\sim}{\beta}}$, the robustness nature of S can be alternatively understood through the first-order Tayor expansion of $f(\underset{\sim}{\beta}_1 \underset{\sim}{Z}')$ around $\underset{\sim}{\beta}_1 = \underset{\sim}{0}$,

$$f(\underset{\sim}{\beta}_1 \underset{\sim}{Z}') \simeq f(0) + \dot{f}(0)\underset{\sim}{\beta}_1 \underset{\sim}{Z}'\ .$$

With the constant terms $f(0)$ and $\dot{f}(0)$ being absorbed into $\underset{\sim}{\beta}_0$ and $\underset{\sim}{\beta}_1$, respectively, the model (3.2) for μ in the small neighborhood of the null hypothesis H_0 can be approximated by the generalized linear model

$$\mu = g(\underset{\sim}{\beta}_0 \underset{\sim}{W}' + \underset{\sim}{\beta}_1 \underset{\sim}{Z}') \ , \tag{3.8}$$

regardless of the form for f. Conversely, the score-test statistic for H_0: $\underset{\sim}{\beta}_1 = \underset{\sim}{0}$ based on (3.8) can be used not only for testing against the linear effect of $\underset{\sim}{Z}$ on $\theta = g^{-1}(\mu)$ but any other monotonic effects, e.g., $f(\underset{\sim}{\beta}_1 \underset{\sim}{Z}') = c\, e^{\underset{\sim}{\beta}_1 \underset{\sim}{Z}'}$ for some unknown c.

4. EXAMPLES

The following examples illustrate how the probability density function h and the inverse "link" function g affect the score-test statistic S of (3.4) through $\underset{\sim}{s}$, D and G given by (3.3) under various situations.

Example 1. The Location Family. Let $y_i = \mu_i + \varepsilon_i$, $i=1,\ldots,n$, where ε_i's are i.i.d. with $E(\varepsilon_i) = 0$ and a known p.d.f. $h(\varepsilon)$. Then, $D^2 = d^2 I$, where $d^2 = E\{-\partial^2 \ell n h/\partial\varepsilon^2\}$ is the intrinsic accuracy of $h(\cdot)$. If $g(t) = t$, i.e., $G = I$, $R(X^*) = R(X)$, so S becomes

$$S = \underset{\sim}{s}\, P_{Z/W}\, \underset{\sim}{s}'/d^2 \ ,$$

with $s(\mu) = H(y-\mu)$, $H(\varepsilon) \equiv -\partial \ell n h/\partial\varepsilon$. For instance, for the student t distribution with f degrees of freedom, $H(\varepsilon) = \text{weight} \cdot \varepsilon$, where the weight $= (f+1)/(f+\varepsilon^2)$ tends to become small as ε becomes large.

Example 2. The Exponential Family. Let $h(y|\mu)$ be of the exponential family,

$$h(y|\mu) = \exp\{A(\mu)y + B(\mu) + D(y)\} \ .$$

Then, by the fact that $s(\mu) = (y-\mu)/\sigma_y^2$, where $\sigma_y^2 = \mathrm{Var}(y|\mu) = 1/d^2(\mu)$, the standardized score vector $\underset{\sim}{e}$ of (3.6) becomes

$$\underset{\sim}{e} = \underset{\sim}{s}\ D^{-1} = (\underset{\sim}{y}-\hat{\underset{\sim}{\mu}})\hat{\Sigma}^{-\frac{1}{2}} , \qquad (4.1)$$

the standardized residual vector, where $\hat{\underset{\sim}{\mu}} = (\hat{\mu}_1,\ldots,\hat{\mu}_n)$ and $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}^2_{y_1},\ldots,\hat{\sigma}^2_{y_n}) = D^{-2}$. With the natural parameter $\theta = A(\mu)$ modelled as $\theta = \underset{\sim}{\beta}_0\underset{\sim}{W}' + f(\underset{\sim}{\beta}_1\underset{\sim}{Z}')$, $g = A^{-1}$ and $G = \hat{\Sigma}$. In this case, $X^* = \hat{\Sigma}^{\frac{1}{2}} X$.

Note that, from (4.1), $\underset{\sim}{e}\ \underset{\sim}{e}' = \Sigma(y_i-\hat{\mu}_i)^2/\hat{\sigma}^2_{y_i} \equiv X^2$ is the Pearson X^2 test statistic for the goodness-of-fit test for the model $\mu = g(\underset{\sim}{\beta}_0\underset{\sim}{W}')$. Hence S of (3.7) can be interpreted as the portion of X^2 due to the $\underset{\sim}{Z}$-effect under H_1. Such tests and their biostatistical applications under the generalized linear models have been discussed by Pregibon (1982), along with the use of the GLIM system for computing S statistics.

Example 3. The F-test. Consider the normal case, $y_i \sim N(\mu_i,\sigma^2)$, $i=1,\ldots,n$, with $\mu_i = \underset{\sim}{\beta}_0\underset{\sim}{W}_i' + f(\underset{\sim}{\beta}_1\underset{\sim}{Z}_i')$. The resulting score-test statistic for H_0 then is

$$S = (\underset{\sim}{y}-\hat{\underset{\sim}{\mu}})\ P_{Z/W}(\underset{\sim}{y}-\hat{\underset{\sim}{\mu}})'/\hat{\sigma}^2 ,$$

where $\hat{\underset{\sim}{\mu}} = P_W\underset{\sim}{y}$ and $\hat{\sigma}^2 = \Sigma(y_i-\hat{\mu}_i)^2/n$. In this case, an equivalently robust test statistic, obtained by substituting ks^2, where $s^2 = \underset{\sim}{y}(I-P_X)\underset{\sim}{y}'/(n-\ell-k)$, for $\hat{\sigma}^2$ in S, is the commonly used F statistic in the linear regression model,

$$F = (\underset{\sim}{y}-\hat{\underset{\sim}{\mu}})\ P_{Z/W}(\underset{\sim}{y}-\hat{\underset{\sim}{\mu}})'/ks^2 ,$$

which has the exact null distribution, F distribution with $(k,\ n-\ell-k)$ degrees of freedom under H_0.

5. FURTHER REMARKS

Remark 1. The immediate extension of the robustness nature of the score tests in the regression context is as follows. Suppose $\underset{\sim}{Z} = (\underset{\sim}{Z}^{(1)}, \ldots, \underset{\sim}{Z}^{(p)})$ with the $\underset{\sim}{Z}$-effect on $\theta = g^{-1}(\mu)$ modelled as $f_1(\underset{\sim}{\beta}_1 \underset{\sim}{Z}^{(1)'}) + \ldots + f_p(\underset{\sim}{\beta}_p \underset{\sim}{Z}^{(p)'})$ under H_1, where f_j's εF. Then the score-test statistic for the null hypothesis of no $\underset{\sim}{Z}$-effect, i.e., H_0: $\underset{\sim}{\beta}_j = 0$, $j=1,\ldots,p$, remains the same regardless of the forms chosen for f_j's. Thus, for instance, the score test for H_0: $\beta_1 = \beta_2 = 0$ for the model $\mu = \beta_0 W + c_1 e^{\beta_1 Z^{(1)}} + c_2 e^{\beta_2 Z^{(2)}}$, where c_1 and c_2 unknown, is the same as that for the linear model $\mu = \beta_0 W + \beta_1 Z^{(1)} + \beta_2 Z^{(2)}$.

Remark 2. If $\mu = g(\beta_0 + f(\underset{\sim}{\beta}_1 \underset{\sim}{Z}'))$, where $\underset{\sim}{W} = 1$, the resulting score test for H_0: $\underset{\sim}{\beta}_1 = \underset{\sim}{0}$ depends neither on f nor on g, and it reduces to the one for the null model considered by Chen (1983).

Remark 3. The reason that the expected Fisher information C is preferred over the observed one, $[-\partial^2 L/\partial\underset{\sim}{\beta}'\partial\underset{\sim}{\beta}]_{\hat{\underset{\sim}{\beta}}}$, in S of (2.1) is mainly because the latter depends on the form of f through its second derivative, which makes it less desirable with regard to the robustness of the test statistic.

ACKNOWLEDGMENTS

I am very grateful to Sanford Weisberg for valuable discussions.

REFERENCES

AITCHISON, J. and SILVEY, S.D. (1958), "Maximum Likelihood Estimation of Parameters Subject to Restraints," Annals of Mathematical Statistics, 29, 813-828.

BREUSCH, T.S. and PAGAN, A.R. (1979), "A Simple Test for Heteroskedasticity and Random Coefficient Variation," Econometrica, 47, 1287-1294.

BREUSCH, T.S. and PAGAN, A.R. (1980), "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics," Review of Economic Studies, 47, 239-254.

BÜHLER, W.J. and PURI, P.S. (1966), "On Optimal Asymptotic Tests of Composite Hypotheses with Several Constraints," Zeitschrift für Wahrscheinlickeitstheorie, 5, 71-88.

CHEN, Chan-Fu (1983), "Score Tests for Regression Models," Journal of the American Statistical Association, 78, 158-161.

COOK, R.D. and WEISBERG, S. (1983), "Diagnostics for Heteroscedasticity in Regression," Biometrika, 70, 1-10.

EICKER, F. (1965), "Limit Theorems for Regressions With Unequal and Dependent Errors," Proceedings of the Fifth Berkeley Symposium I, 59-66.

MORAN, P.A.P. (1970), "On Asymptotically Optimal Tests for Composite Hypotheses," Biometrika, 57, 47-55.

MORAN, P.A.P. (1973), "Asymptotic Properties of Homogeneity Tests," Biometrika, 60, 79-85.

NELDER, J.A. and WEDDERBURN, R.W.M. (1972), "Generalized Linear Models," Journal of Royal Statistical Society, A, 135, 370-384.

NEYMAN, J. (1959), "Optimal Asymptotic Tests for Composite Hypotheses," in Probability and Statistics, ed. U. Grenander, New York: John Wiley, 213-234.

NEYMAN, J. and SCOTT, E. (1966), "On the Use of C(α) Optimal Tests of Composite Hypotheses," Bulletin de l'Institut International de Statistique, 41, 477-497.

PREGIBON, D. (1982), "Score Tests in GLIM With Applications," in Lecture Notes in Statistics, no. 14, GLIM.82: Proceedings of the International Conference on Generalized Linear Models (ed. R. Gilchrist), Spring-Verlag, New York.

RAO, C.R. (1948), "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters With Applications to Problems of Estimation," Proceedings of the Cambridge Philosophical Society, 44, 50-57.

TARONE, R.E. (1982), "The Use of Historical Control Information in Testing for a Trend in Poisson Means," Biometrica, 38, 457-463.

TARONE, R.E. and GART, J.J. (1980), "On the Robustness of Combined Tests for Trends in Proportions," Journal of the American Statistical Association, 75, 110-116.

WEISBERG, S. (1983), "Discussion of Developments in Linear Regression Methodology: 1959-1982," by R.R. Hocking, Technometrics, 25, 240-244.